DOCUMENT RESUME

ED 138 630 TM 006 253

AUTHOR Meredith, John B., Jr.
TITLE The Use of a Criterion-Referenced Test Methodology to Assess Training System Effectiveness.
PUB DATE [Apr 77]
NOTE 7p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Analysis of Variance; *Criterion Referenced Tests; *Cutting Scores; Item Analysis; *Military Training

ABSTRACT

Criterion referenced test methodology is used in this study for determining the extent of overtraining and undertraining through the use of the minimally acceptable performance level (MAPL), as proposed by Nedelsky (1954). A MAPL was computed for each area on a 320-item, 12-area test. The raw area scores of the technicians were transformed to standardized area criterion scores. An analysis of variance was performed using a fixed effects model with paygrade categories nested under each area. The null hypothesis was rejected at the .005 level of significance. A post-hoc multiple comparison (Newman-Keuls) was used to determine which areas indicated the possibilities of overtraining or undertraining. The results were verified by subject matter experts. (Author/RC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED138630

THE USE OF A CRITERION-REFERENCED TEST METHODOLOGY TO ASSESS TRAINING SYSTEM EFFECTIVENESS

by

John B. Meredith, Jr.
Data-Design Laboratories
Norfolk, Virginia 23502

U S DEPARTMENT OF HEALTH EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

INTRODUCTION

In developing criterion-referenced tests, one of the most difficult problems is the determination of appropriate passing scores. Historically, passing scores have been set at an arbitrary level between 80% and 90% correct (Sanders, 1976; Shaycoft, 1976). The use of arbitrary passing scores has two major disadvantages: classification error (Reichman and Oosterhof 1976), and nonconsideration of item difficulty. In this study, a technique originally proposed by Nedelsky (1954) was used for the determination of minimally acceptable performance levels (MAPL) for a group of Navy technician trainees. The analysis of the proportions of technicians achieving the MAPL in various subject matter areas was used as a basis for assessing training system effectiveness for each area.

Alteration of the passing score can be used as a means to make personnel classifications more or less definitive. One of the most logical methods for altering the passing score is to consider item difficulty as a measure of the similarily between the distracters and the correct response. In fact, this approach to the problem of setting the passing score yields the best estimate of the probability of classifying an individual into the correct mastery or nonmastery state (Reichman and Oosterhof, 1976). Item difficulty, in this sense, is criterion-referenced since it relates to test content rather than to performance relative to a specific group of personnel (Smilansky and Guerin, 1976).

The Nedelsky procedure has been refined and supported by Taylor and Reid (1972), Bobula (1974), Meredith and Robinson (1976), and Smilansky and Guerin (1976). The use of this procedure is dependent upon the assumption that subject matter experts (SMEs) can define the similarity between distracters and the correct response for each test item, independent of the actual testing situation.

The value of item similarity is assigned as follows:

a. an alternative which a minimally qualified person should recognize as incorrect is assigned a value of zero,

b. an alternative which a minimally qualified person should not recognize as incorrect is assigned a value of two,

TM006 253

c. an alternative which is the correct response is assigned a value of two, and

d. an alternative which does not meet any of the above criteria is assigned a value of one.

An example of the application of this MAPL technique, adapted from Bobula (1974), is exemplified by the faculty member who is teaching statistics and defines ability to recognize measures of central tendency as a component of basic statistical competency. He might write an item and assign the following values:

	Value	Item
		"The most appropriate measure of central tendency for the reading performance scores of a large class is the"
	(2)	1. Mo
	(0)	2. V ce
Correct	(2)	3. Me
	(2)	4. Mea n
	(0)	5. Standard deviation

When the SMEs rated this item for the minimally acceptable student they decided that three alternatives (Mean, Median, and Mode) were equally viable, and all were given values of two (2).

The Alternative Similarity Index (ASI) of an item is defined as follows:

$$ASI = \frac{\text{Value of correct alternative}}{\text{Sum of values of all alternatives}} = \frac{2}{6} = .33$$

The ASI of each item is the probability that a minimally qualified person would select the correct alternative by chance (Smilansky, 1976). The rounded sum of the ASIs in each test area is defined as the MAPL for that test area. A passing score is given when the technician's score is equal to or greater than the MAPL. The percentage of technicians achieving the MAPL is determined by paygrade for each area of the test.

From standardized area criterion scores of the technicians, an analysis of variance design is used to test differences among training areas. Any area of the test that is significantly below the other areas indicates the possibility of undertraining. Conversely, any area of the test that is significantly above the other areas indicates the possibility of overtraining. Additional SMEs were used to validate any indications of overtraining and/or undertraining.

METHOD AND RESULTS

In this study, five SMEs evaluated a 360-item test to determine the MAPL for each of 12 areas of specialization. The area MAPLs are shown in Table 1.

Table 1. Area MAPLs

Area	No. of Items	MAPL
A	10	6
B	29	14
C	30	15
D	46	25
E	32	16
F	46	24
G	24	11
H	33	15
I	24	11
J	29	15
K	32	15
L	25	14

Using each area MAPL, the percentage of technicians (n=274) achieving the MAPL, by paygrade category is presented in Table 2.

Table 2. Percentage of Technicians, by Paygrade Category, Achieving Each Area MAPL

Area	E3/4	E5	E6	E7/8
A	56.4	91.0	100.0	100.0
B	90.0	98.0	95.3	100.0
C	57.3	92.0	95.3	100.0
D	51.0	84.0	95.3	100.0
E	63.6	74.0	86.0	95.2
F	39.1	73.0	95.3	90.5
G	50.0	77.0	93.0	95.2
H	49.1	73.0	88.4	95.2
I	65.5	84.0	100.0	95.2
J	32.7	35.0	62.8	71.4
K	38.2	62.0	86.0	90.5
L	44.5	64.0	69.8	85.7

An analysis of variance was performed using a fixed effects model with the four paygrade categories nested under each area. The raw area scores of the technicians were transformed to a standardized area criterion score, Z_i, using the following:

$$Z_i = \frac{X_i - MAPL_i}{S_i}$$

Where: X_i = raw test score for i^{th} area,

$MAPL_i$ = criterion score for i^{th} area (See Table 1.)

S_i = standard deviation of raw test score for i^{th} area.

The null hypothesis was that there would be no significant difference among areas.

$$Ho: \mu_a = \mu_b = \ldots = \mu_l$$

The null hypothesis was rejected at the .005 level of significance. The analysis of variance summary is presented in Table 3.

Table 3. Analysis of Variance Summary

Source of Variation	Sums of Squares	Degrees of Freedom	Mean Square	F
Among Areas	768.73	11	69.88	5.29*
Among Paygrades	476.04	36	13.22	
Error	2,529.03	3240	.78	
Total	3,773.80	3287		

*Significant at $p < .005$ with 11,36 df

A post-hoc multiple comparison test (Newman-Keuls) was performed to determine which standardized area criterion scores differed significantly from the others. It was found that Area B was significantly higher and that Area J was significantly lower than the other areas ($p < .01$). These results suggest the possibility of overtraining in Area B and undertraining in Area J.

The validity of these indications of overtraining and undertraining findings was confirmed by additional SMEs, independent of the above analysis. It was found that these SMEs had recommended that training in Area B should be reduced from 7 to 5 days. It was also confirmed that personnel in Area J were undertrained at this time due to the addition of new and unfamiliar equipment.

CONCLUSION

This paper has presented a methodology for making preliminary training assessments from criterion-referenced, multiple-choice tests. An innovative technique for obtaining criterion level information was described and applied to the evaluation of test results. Using these assessments, the relative emphasis in the various training areas can be modified to better attune personnel to the operational requirements of the system.

REFERENCES

1. Bobula, J. Minimum Pass Level Study. Report to the Faculty, Center for Educational Development, College of Medicine, University of Illinois, 25-31, 1974.

2. Meredith, J. B., and Robinson, M. A., Determination of Minimal Acceptable Performance Levels for Criterion-Referenced, Multiple-Choice Tests. Paper presented at the 18th Annual Conference of the Military Testing Association. Gulf Shores, Alabama, October, 1976.

3. Meskauskas, J. A. Evaluation Models for Criterion-Referenced Testing: Views Regarding Mastery and Standard Setting, Review of Educational Research, 46:133-158, 1976.

4. Nedelsky, L. Absolute Grading Standards for Objective Tests. Educational and Psychological Measurement, 14:3-19, 1954.

5. Reichman, S. L. & Oosterhof, A. C. Strategy Guidelines for the Construction of Mastery Tests, Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California, April, 1976.

6. Sanders, J. R. Alternatives for Achievement Testing, Educational Technology, 16:March, 1976.

7. Shaycoft, M. F. A Paradox in Setting Cutting Scores on Criterion-References Tests, Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California, April, 1976.

8. Smilansky, J. and Guerin, R. O. Minimal Acceptable Performaance Levels for Criterion-Referenced Multiple-Choice Examinations and Their Validation, Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, California, April 1976.

9. Taylor, D. D. & Reid, J. C. Criterion-Referenced Evaluation of Student Performance, Journal of Medical Education, 47:970-971, 1972.